

CHAPTER 3—TEST DEVELOPMENT PROCESS

DEVELOPMENT COMMITTEE ITEM IDEA GENERATION

The development of the MEA tests continued to be a cooperative effort by committees of Maine teachers, curriculum supervisors, higher education faculty, content specialists of the Department of Education, and curriculum/assessment specialists employed by the program's contractor, Advanced Systems in Measurement and Evaluation, Inc. The committees were structured to represent all areas of the state and committee members all served rotating terms. One of the Appendices contains a listing of MEA development committee members for 1998–1999.

The committees' primary roles were to develop test questions for the MEA and to interpret testing data so that those questions could be selected for the program. The MEA development committee for each subject area at grade levels 4, 8, and 11 met several times. In the development phase, the committees reviewed the content standards and test specifications; they brainstormed or drafted test questions and scoring rubrics to fit those specifications. After the questions were field tested, the committees reviewed the field-test data and made recommendations about selecting, revising, or eliminating specific questions from the item pool for the operational test. At that time, the committees also confirmed that each question conformed directly to Maine's *Learning Results* and was thus assigned to the appropriate content standard reported in school and district results. Because many MEA questions are released to the public each year, the committees repeat these activities annually, as new questions are developed in order to replenish the item pool.

INTERNAL ITEM REVIEW

- The lead or peer test developer within the content specialty reviewed the typed item, open-response scoring guide, and any reading selections and graphics.

- The content reviewer considered item “integrity;” item content and structure; appropriateness to designated content area; item format; clarity; possible ambiguity; keyability; single “keyness;” appropriateness and quality of reading selections and graphics; and appropriateness of scoring guide descriptions and distinctions (as correlated to the item and within the guide itself).
- The content reviewer also considered scorability and evaluated whether the scoring guide adequately addressed performance on the item.
- Fundamental questions the content reviewer considered, but was not limited to, included the following:
 - What is the item asking?
 - Is the key the only possible key?
 - Is the open-response item scorable as written (were the correct words used to elicit the response defined by the guide)?
 - Is the wording of the scoring guide appropriate and parallel to the item wording?
 - Is the item complete (e.g., with scoring guide, content codes, key, grade level, and contract identified)?
 - Is the item appropriate for the designated grade level?

EXTERNAL ITEM REVIEW

- Item sets were brought to Development Advisory Committee meetings for review and revision.

ITEM EDITING

Editors reviewed and edited the items from the Development Advisory Committee item review to ensure uniform style (based on *The Chicago Manual of Style, 14th Edition*) and adherence to sound testing principals.

These principals included the stipulation that items

- were correct with regard to grammar, punctuation, usage, and spelling;
- were written in a clear, concise style;
- contained unambiguous explanations to students as to what is required to attain a maximum score;

- were written at a reading level that would allow the student to demonstrate his or her knowledge of the tested subject matter, regardless of reading ability;
- exhibited high technical quality regarding psychometric characteristics;
- had appropriate answer options or score-point descriptors; and
- were free of potentially insensitive content.

REVIEWING AND REFINING

Test developers presented item statistics to the development committees to assist in the committees' recommendation for placement of items into the common and matrix portions of the test. The Department of Education made the final selections with the assistance of Advanced Systems at a meeting.

OPERATIONAL TEST ASSEMBLY

Test assembly is the sorting and laying out of item sets into test forms. Criteria considered during this process included the following:

- Content coverage/match to test design. The curriculum specialist completed an initial sorting of items into sets based on a balance of content categories across sessions and forms, as well as a match to the test design (e.g., number of multiple-choice, short-answer, and open-response items).
- Item difficulty and complexity. Item statistics drawn from the data analysis of previously tested items were used to ensure that there were similar levels of difficulty and complexity across forms.
- Visual balance. Item sets were reviewed to ensure that each reflected a similar length and “density” of selected items (e.g., length/complexity of reading selections, or number of graphics).
- Option balance. Each item set was checked to verify that it contained a roughly equivalent number of key options (As, Bs, Cs, and Ds).
- Name balance. Item sets were reviewed to ensure that a diversity of names was used.

- Bias. Each item set was reviewed to ensure fairness and balance based on gender, ethnicity, religion, socio-economic status, and other factors.
- Page fit. Item placement was modified to ensure the best fit and arrangement of items on any given page.
- Facing page issues. For multiple items associated with a single stimulus (a graphic or reading selection), consideration was given to whether those items needed to begin on a left- or right-hand page, as well as to the nature and amount of material that needed to be placed on facing pages. These considerations served to minimize the amount of “page flipping” required of the students.
- Relationships between forms. Sets of “common” items were placed identically in each version of the forms. Although matrix-sampled item sets differ from form to form, they must take up the same number of pages in each form so that sessions and content areas begin on the same page in every form. Therefore, the number of pages needed for the longest form often determines the layout of each form.
- Visual appeal. The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of “white space,” the density of the text, and the number of graphics.

EDITING DRAFTS OF OPERATIONAL TESTS

Any changes made by the test construction specialist must be reviewed and approved by the test developer. Once a form had been laid out in what was considered its final form, it was reread to identify any final considerations, including the following:

- Editorial changes. All text was scrutinized for editorial accuracy, including consistency of instructional language, grammar, spelling, punctuation, and layout. Advanced Systems’ publishing standards are based on *The Chicago Manual of Style, 14th Edition*.

- “Keying” items. Items were reviewed for any information that might “key” or provide information that would help answer another item. Decisions about moving keying items are based on the severity of the “key-in” and the placement of the items in relation to each other within the form.
- Key patterns. The final sequence of keys was reviewed to ensure that their order appeared random (e.g., no recognizable pattern, and no more than three of the same key in a row).

BRAILLE AND LARGE-PRINT TRANSLATION

Form one for grades 4, 8, and 11 tests was translated into Braille by a subcontractor who specializes in test materials for blind and visually-handicapped students. In addition, form one for each grade was adapted into a large-print version.